DOCUMENT RESUME

ED 297 305 CS 009 260

AUTHOR Perkins, Kyle; Parish, Charles
TITLE What's Wrong with Reading Comprehension Tests?
PUB DATE 88
NOTE 18p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cognitive Processes; *English (Second Language); Multiple Choice Tests; *Prior Learning; *Reading Comprehension; Reading Research; *Test Validity
IDENTIFIERS Test of English as a Foreign Language; Text Factors

ABSTRACT

This document reviews studies that investigated certain inadequacies of English as a Second Language (ESL) reading comprehension tests. The first study examined whether three ESL reading comprehension tests required information principally from the text or from background knowledge. Results indicated that background knowledge differentially affects reading comprehension at different levels of proficiency and that a large proportion of the variance in reading comprehension scores can be accounted for by the reader's background knowledge. The second study measured reasoning in the Test of English as a Foreign Language (TOEFL) reading comprehension subtest. Results indicated that the subtest seemingly lacks validity. The third study investigated whether an ESL test distinguished between the subskills it professed to measure. Results indicated that only 44% of the total test exhibited internal construct validity. The fourth study examined behavioral anchoring, whose purpose is to define achievement and to make test scores interpretable in terms of what students at different ability levels could and could not do. Results indicated that higher ability level students excelled with items which depended on implicit information; that all proficiency levels exhibited achievement with linguistic structures which related parts of the text to one another; and that the derivational complexity and the readability level estimates co-varied with an increase in proficiency level. Specific remedies for each of the weaknesses of ESL reading comprehension tests are suggested. (Twenty-six references are appended.) (RS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Kyle Perkins
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

What's Wrong With Reading Comprehension Tests?

Kyle Perkins & Charles Parish

Southern Illinois University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED297305

Our purpose in this paper is to present our perception of certain inadequacies in ESL reading comprehension tests, among which are the following: 1) the common insistence on a single correct answer; 2) their lack of validity if they do not measure the ability to apply reasoning; 3) their failure to distinguish between the putative subskills they purport to measure; and 4) the general lack of behaviorial anchoring descriptions.

The insistence on a single correct answer

Many ESL reading comprehension tests are the multiple-choice type with a single correct keyed response. Such tests may inadvertently penalize the creative reader who is capable of proceeding beyond the conventional, expected implications of a text and choosing an alternative that is different from the anticipated and allowed keyed response.

Creative readers may also be trapped into choosing a non-allowed foil as the correct response for a reading comprehension item, especially an inferenced-based item, because of the nature of their particular stocks of textually-relevent background knowledge. As Farr and Carey (1986) have noted, in their summary of the importance of background knowledge, "meaning is as dependent on the reader as it is on the text" (p. 34). The point that we want to make here is that different amounts and different levels of richness of textually-relevant background knowledge can lead the more creative readers astray in terms of ticking the correct response.

We have alluded to the importance of background knowledge in the reading comprehension process and we offer the following observation of its importance:

AS009260

> The emerging model of comprehension asserts that comprehension is an active process in which the reader constructs meaning from text cues, calling upon knowledge of language, text structure and conventions, content concepts and communication. This process is essentially inferential with readers using their existing knowledge to link discrete pieces of information in the text, to ascribe appropriate meanings to words, and to fill in implied information. In addition, readers monitor their state of understanding and engage various strategies according to different purposes and genre, or to encountered difficulties (Farr, Carey, and Tone, 1986, p. 136).

Because determining the influence of a reader's textually-relevant background knowledge, and consequently, determining the passage dependency of reading comprehension test items, are crucial in reading assessment, reading researchers constantly seek to employ techniques which have been proposed to identify the effect of background knowledge on reading comprehension test items. Perkins and Brutten (1988) employed Schlesinger and Weiser's (1970) facet design to determine whether the items in three ESL reading comprehensions required information principally from the text or from the reader's textually-relevant background knowledge. Three facets or structs (the form of information [explicit, implicit]; the source of the information [text, nontext]; and the frequency with which the information appeared in the text) were entered into regression equations to determine how much variance the structs accounted for in the item scores from ESL reading comprehension tests from different levels of proficiency.

The results were as follows. For the elementary proficiency level, no facet accounted for significant variance. For the intermediate-level test, frequency and source accounted for significant variance in the criterion, and the regression model accounted for 52.85 percent of variance in the criterion. For the advanced-level test, the variables entered

the equation in the following order: frequency, source, form, and the total regression model accounted for 43.71 percent of the variance in the criterion.

The study concluded that 1) the effects of the facets (including background knowledge) differentially affect reading comprehension at different levels of proficiency, and 2) as many L1 studies have shown, a large proporition of the variance in a measure of L2 reading comprehension can be accounted for by the reader's prior knowledge of the content matter contained in the prose passages that are tested.

Spiro and Meyers (1984) have catalogued a list of factors which can account for individual differences in readers' answers to questions vis-à-vis textually-relevant background knowledge: (1) some readers may generally lack background knowledge; (2) some readers may under-utilize background knowledge; (3) others may over-utilize background knowledge; (4) some readers may have difficulty accessing background knowledge; (5) others may experience difficulty determining relevant background knowledge; (6) there may be a mismatch between the nature of the text and background knowledge; and (7) lack of background knowledge may cause a short-term memory overload of keeping the stem question or stimulus in mind as the reader processes the alternatives.

In sum, given the fact that the creative reader may go beyond what is anticipated and what is expected for a correct answer and that readers demonstrate different degrees of facility with background knowledge, we come full circle to an axiom that has long been acknowledged by test constructors: the grading of a multiple-choice test is quite objective, but the determination of the "single, best correct answer" to a reading comprehension test item may be quite subjective, appropriately varying according to the reader's store of textually-relevant background knowledge.

Applying reasoning in reading

Many reading researchers have attached great importance to the role of reasoning in reading. For example, Thorndike (1917) concluded that reading comprehension involved the application of thinking skills to printed materials; and he likened understanding a paragraph to solving a problem, noting that the reader's mind must "select, repress, soften, emphasize, correlate and organize" during reading comprehension. In an attempt to isolate specific reading skills, Davis (1944) submitted the results of nine reading comprehension tests administered to a group of college freshman to factor-analytic techniques and found that word knowledge and reasoning accounted for 89 percent of the total variance. Bruner's (1957) observation that the mind uses existing information to interpret and to organize new information in abstractions, and relations not directly given by the text have served as the foundation for understanding reading as an inferential or reasoning process. Comprehension is an interactive, inferential process whereby readers construct meaning from the text by using their existing knowledge to connect pieces of information in the text and to supply implied information (Bransford, Barclay and Franks, 1972). Baker and Brown (1984) noted that "an important element of comprehension is understanding how and why ideas in a text are interconnected" (p. 31).

Although there is a well-established tradition of investigating reason in native-language reading comprehension, the measuring of reasoning in second language reading comprehension is relatively new. Perkins (1988) reported the results of an attempt to measure the ability of adult ESL students to apply reasoning in reading. The issue was to determine whether the TOEFL reading comprehension test measures reasoning in reading. The TOEFL was chosen as the criterion measure because it is a well-

known and widely-used measure of ESL proficiency.

The experimental measure was Kretschmer's (1972) test which assesses subjects' application of thinking skills to printed text. Kretschmer's test was developed from a model of reading comprehension based upon cognitive activities which Piaget (1967) referred to as "concrete operations." In the Piagetian paradigm, concrete operations are so called because they were based on activities using concrete props. Kretschmer delineated Piaget's concrete operations in prose text, his test establishing the following operations: (1) an additive classification, (2) a multiplicative classification, (3) an additive series, and (4) a multiplicative series. Because Kretschmer's test was designed to measure subjects' application of thinking skills or reasoning to printed text, it is clearly a measure of cognitive ability applied to a reading medium.

A point-biserial correlation was calculated between each reasoning in reading item and the total scores for the reasoning in reading test and then corrected for part-whole overlap. Another point-biserial correlation was calculated between each reasoning in reading item and the TOEFL reading comprehension total scores. If a reasoning in reading item was more highly correlated with the reasoning in reading total than with the TOEFL reading comprehension total, that item was declared to have shown construct validity. Only twenty percent of the TOEFL reading comprehension items had higher correlations with the reasoning in reading test than with the TOEFL reading comprehension total scores. The conclusion was that Kretschmer's experimental test measures a cognitive ability applied to a reading medium, reasoning in reading, and that the TOEFL reading comprehension subtest seemingly lacks validity because it does not share a substantial amount of trait variance with the reasoning

test. This position has merit if one accepts Thorndike's insight that reading comprehension involves the application of thinking skills to printed text.

Failure to distinguish between putative subskills

First-language reading comprehension researchers have been concerned with whether reading assessment instruments distinguish between the subskills that they profess to measure; however, little research of this nature has been conducted on ESL reading measures. Perkins and Parish (1988) reported the results of a construct validity study of items intended to measure three major skills objectives in an ESL reading comprehension test. The three major skills objectives were (1) facts, (2) generalizations, and (3) inferences. The task was to determine whether the fact questions measured a construct different from the constructs measured by the generalization questions and by the inference questions. In other words, could evidence be found to show that the sets of questions designed to measure the three different major skills objectives did indeed measure something unique?

Perkins and Parish used the skills objectives definitions found in the Preliminary Teacher's Guide for the Iowa Tests of Basic Skills to categorize the 25 questions of the multiple-choice type with a single correct keyed response from an advanced-level ESL reading comprehension test. Those definitions are "Facts: to recognize and understand stated factual details and relationships (literal meaning); Generalizations: to develop generalizations from a selection (evaluative meaning); Inferences: to infer underlying relationships (interpretative meaning)" (Hieronymus, Hoover, and Lindquist, 1986, p. 45).

For each item a point-biserial correlation was calculated between

its own subtest total score and the total scores of the two other subtests; for purposes of analysis, the total test was partitioned into three subtests, (1) items assessing facts, (2) items assessing generalizations, and (3) items assessing inferences.

Perkins and Parish utilized Henning's (1987) internal construct validation procedure, which, in the context of this research, makes the following assumptions. If, for example, the fact items have construct validity, then the point-biserial correlation between each fact item and the fact subtest total scores (corrected for part-whole overlap) should be higher than the point-biserial correlations of the fact items with the generalization subtest total scores or with the inference subtest total scores.

For each set of questions designed to measure a major skill objective, the authors determined the proportion of valid items within that subtest. Valid in this context means that an item correlated more highly with its own subtest total than with the total from another subtest.

The results, which suggest an opportunity for economy of measurement, were the following. There were only three fact items which had fact item with fact total correlations greater than both the fact-generalization total and fact-inference total correlations. There were only two generalization items which had generalization-generalization total correlations greater than the generalization-fact total correlations and the generalization-inference total correlations. And there were six inference items that had inference-inference total correlations greater than the inference-fact total correlations and the inference-generalization total correlations.

In sum, only 11 of the 25 items (or 44 percent of the total test) exhibited internal construct validity when they were compared with both

criterion measures. If the purpose of measurement is to assess how well students recognize facts, then only three fact items measured something uniquely factual. It does not seem warranted to use a fact item to elicit information about how well students recognize and comprehend factual information when that particular item correlates higher with the inference total than it correlates with its own fact subtest total. Thus, little evidence could be marshalled from this ESL test to show that the three different tasks reflected three different and independent constructs.

Behavioral anchoring analysis

Until very recently there were no means to provide a description of what students at different points on a proficiency continuum could do that students at lower points on the ability continuum could not do. Fortunately, this inadequacy has been accommodated by the introduction of behavioral anchoring, a procedure recently developed by the National Assessment of Educational Progress (Messick, Beaton, and Lord, 1983; National Assessment of Educational Progress, 1985). The purpose of behavioral anchoring is to define achievement and to make test scores interpretable in terms of what students at different ability levels could and could not do.

Although the National Assessment of Educational Progress has been issuing behavioral anchoring analyses for L1 elementary and secondary educational achievement for nearly ten years, no such analyses have been conducted with L2 data sets. To supply some benchmark L2 research, Perkins and Brutten (1988b) prepared behavioral anchoring analyses for three ESL reading comprehension tests following the general procedures for behavioral anchoring detailed in Beaton (1985). First, three anchor points were selected for each test where items could be found that discriminated between what students at each point could do that students

at lower points on the ability scale could not do. Next, the items found to discriminate between pairs of anchor points were batched, and the items and the corresponding texts from which the items were derived were analyzed. From that analysis, a description of what the items at each level represented was written. The descriptions were based on three foci: (1) the relationship of each item to the structure of the test, (2) the reader's prior knowledge, and (3) the nature of the requisite cognitive processes entailed in answering the question.

Meyer's (1975) prose analysis system was used as a guideline for a discussion of how the items related to the structure of the text. Meyer's system utilized a detailed outline or tree structure which she termed a content structure, and it relates a text's information and shows how some textual ideas are superordinate to other ideas. There are three distinct levels of content structure. The top-level structure (the top-third level) is the rhetorical relationship that binds together all the propositions in a text, giving it texture, and it is the part of the content structure which contains the main ideas, termed macropropositions. The middle third of the total content structure contains the major supporting details that complement the major premises of the text's message, and these entities are called micropropositions. The bottom third of the total content structure contains very specific, minor details.

To ascertain the extent to which the correct responses to the discriminating items were dependent on the reader's prior knowledge, the authors used Schlesinger and Weiser's (1970) facet design and Pearson and Johnson's (1978) three-part classification scheme for reading comprehension items, which focuses on the relations between each question and its keyed response. Pearson and Johnson's scheme entails the following

comprehension categories: textually explicit, textually implicit, and scriptally implicit question-answer relations.

Carpenter and Just (1986) provided the framework for a consideration of cognitive processes for the behavioral anchoring study. They discussed two issues which seemed particularly germane: (1) "language processes: syntax and semantics" and (2) specific linguistic structures, such as pronouns, repeated references, and connectives.

"Language processes: syntax and semantics" refers to the effects that the structural characteristics of language have, in particular, the characteristics of sentences that have an effect on the reading process. Perkins and Brutten were particularly interested in the derivational complexity of the texts' sentences which were related to the keyed responses of the discriminating items.

The results of the behavioral anchoring analyses of the three ESL reading comprehension tests were the following. First, for each proficiency level, the higher-ability subjects showed competence with items that dealt with micropropositions which are found in the lower two-thirds of a Meyer total content structure analysis, while the lower-ability subjects showed competence with items that dealt with the macropropositions, which are found in the upper one-third of a total content structure analysis. Second, for each proficiency level, the higher-ability level students excelled with items whose keyed responses depended on implicit sources of information and which appeared at a lower frequency; the lower-ability subjects did not do this. Third, all proficiency levels exhibited achievement with linguistic structures which related parts of the text to one another. And fourth, the derivational complexity and the readability level estimates co-varied with an increase in proficiency level.

Perkins and Brutten reported that the following advantages accrue from a behavioral anchoring analysis. First, a common scale derived from Item Response Theory was employed which permitted performance to be compared across different ability levels within a proficiency-level group. With a very large data set, a common scale could be used to compare performances across different language groups. Second, behavioral anchoring allows the researcher to relate group performance to scales which are anchored by clusters of benchmark items. These clusters of benchmark items provide concrete examples with associated real-world reading capabilities which are, in turn, used for deriving verbal summaries of reading achievement at each scale level. Third, with the Item Response Theory test statistics, it is possible to estimate at any point on the ability scale the proportion of students who can reasonably be expected to pass or to fail any given item. This type of information could be used to establish new criterion-setting procedures or to inform established standard-setting procedures. Fourth, it is possible to extend the utility of behavioral anchoring data by linking them to cognitive, language background, and educational variables in order to further inform second language acquisition research. And fifth, the use of Item Response Theory models generates invariance of item difficulty and person ability across sets of items, something which is not possible with norm-referenced or criterion-referenced measurement.

Varia

Finally, there are additional, amorphous problems with reading comprehension tests which do not lend themselves to any generic classification. First, assessment is intended to be a systematic sampling of reading behaviors which is conducted for the purpose of making a decision or a state-

ment about the readers' ability to read and comprehend. The test constructor must select a relatively small sample of questions from the universe of all possible questions, and therein lies the difficulty. Drum, Calfee, and Cook (1980) have noted that reading comprehension questions differ widely in their information-processing requirements. A second difficulty in item construction and selection is that certain information in a test is more important than other information, e.g. macropropositions versus micropropositions. A logical conclusion that one could draw is that a random sampling of questions may not generate very effective information about reading comprehension.

We end our presentation with the following observation from Farr and Carey (1986):

> Most multiple choice reading comprehension tests reflect grossly inadequate theories of cognition, language, and learning. Certainly, they reflect outdated theories of the reading process. There is no consensus regarding a definition of comprehension, but there are few reading theorists today who would condone a definition of comprehension as word based, a function of short-term recall, or unidimensional (p. 35).

Recommendations

We would be remiss if we did not attempt to suggest remedies for the weaknesses in reading comprehension assessment which we have just reviewed. With regard to the insistence on a single correct answer, perhaps for each multiple-choice question we could ask the students why they chose the answer they did and conduct a protocol analysis to determine if the more creative readers were led astray, and why. With regard to the paucity of reasoning in reading, we suggest that more use be made of the general classes of inferences or reasoning that could be made from a single sentence. Dehn (1984) listed some sixteen classes of inferences: specification, causative, resultative, motivational, enablement, function, enablement-prediction,

missing enablement, intervention, action prediction, knowledge-propagation, normative, state duration, situation, feature, and utterance-intent.

With regard to the failure of items to distinguish between the putative subskills that they purport to measure, we suggest that researchers examine their assessment instruments to determine whether different tasks reflect different independent constructs. If the answer is no, then we suggest economy of measurement by reducing the redundant items in the tests.

And finally, we suggest the use of behavioral anchoring or Haertel's (1987) binary skills model of structuring item domains, procedures which cluster items to form a scale to the extent that readers who are able to answer items in a given cluster could also answer items in clusters lower on the scale. These approaches allow us to address the fact that a random sample of items may differ in information-processing requirements and varying text difficulty from which the items were derived.

REFERENCES

Baker, L., & Brown, A. L. 1984. Cognitive monitoring in reading. In J. Flood (Ed.), Understanding reading comprehension. Newark, Delaware: International Reading Association.

Beaton, A. 1985. Implementing the new design: The NAEP technical report. Report No. 15-TR-20. Princeton, New Jersey: National Assessment of Educational Progress, Educational Testing Service.

Bransford, J.D., Barclay, J., & Franks, J. J. 1972. Sentence memory: constructive vs. interpretive approach. Cognitive Psychology, 3, 193-209.

Bruner, J. S. 1957. Going beyond the information given. In H. Gruber et al. (Eds.), Contemporary approaches to cognition. Cambridge, Massachusetts: Harvard University Press.

Carpenter, P. A., & Just, M. A. 1986. Cognitive processes in reading. In J. Orasanu (Ed.), Reading comprehension: from research to practice. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Davis, F. B. 1944. Fundamental factors of comprehension in reading. Psychometrika, 9, 185-197.

Dehn, N. 1984. An AI perspective on reading comprehension. In J. Flood (Ed.), Understanding reading comprehension. Newark, Delaware: International Reading Association.

Drum, P. A., Calfee, R. C., & Cook, L. K. 1980. The effects of surface structure variables on comprehension in reading comprehension tests. Reading Research Quarterly, 16, 486-513.

Farr, R., & Carey, R. F. 1986. Reading: What can be measured?, second edition. Newark, Delaware: International Reading Association.

Farr, R., Carey, R. F., & Tone, B. 1986. Recent theory and research into the reading process: implications for reading assessment. In J. Orasanu (Ed.), Reading comprehension: from research to practice. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Haertel, E. 1987, April. Structuring item domains to map the school curriculum. Paper presented at the meeting of the American Educational Research Association, Washington, D.C.

Henning, G. H. 1987. A guide to language testing. Cambridge, Massachusetts: Newbury House Publishers.

Hieronymus, A. N., Hoover, H. D., & Lindquist, E. F. 1986. Preliminary teacher's guide. Chicago: The Riverside Publishing Company.

Kretschmer, J. C. 1972. Measuring reading comprehension (72-27,045). Ann Arbor, Michigan: University Microfilms International.

Messick, S., Beaton, A., & Lord, F. 1983. National Assessment of Educational Progress reconsidered: a new design for a new era. NAEP Report 83-1. Princeton, New Jersey: National Assessment of Educational Progress, Educational Testing Service.

Meyer, B. J. F. 1975. The organization of prose and its effect on memory. Amsterdam: North-Holland Publishing Company.

National Assessment of Educational Progress. 1985. The reading report card. ETS Report No. 15-R-01. Princeton, New Jersey: Author, Educational Testing Service.

Pearson, P. D., & Johnson, D. D. 1978. Teaching reading comprehension. New York: Holt, Rinehart and Winston.

Perkins, K. 1988. Measuring ESL readers' ability to apply reasoning in reading: a validity study of the TOEFL reading comprehension subtest. Journal of Research in Reading, 11, 36-49.

Perkins, K., & Brutten, S. R. 1988a. Using a facet design to assess the effects of the form, the source, and the frequency of the information necessary to answer ESL reading comprehension questions. Unpublished manuscript.

Perkins, K., & Brutten, S. R. 1988b. A behavioral anchoring analysis of three ESL reading comprehension tests. Unpublished manuscript.

Perkins, K., & Parish, C. 1988, April. An internal construct validity study of three ESL reading comprehension tests. Paper presented at the 38th Southeastern Conference on Linguistics meeting, Memphis, Tennessee.

Piaget, J. 1967. Six psychological studies. New York: Vintage Books.

Schlesinger, I. M., & Weiser, Z. 1970. A facet design for tests of reading comprehension. Reading Research Quarterly, 5, 566-580.

Spiro, R. J., & Myers, A. 1984. Individual differences and underlying cognitive processes in reading. In P. D. Pearson (Ed.), Handbook of reading research. New York and London: Longman.

Thorndike, E. L. 1917. Reading as reasoning; a study of mistakes in paragraph reading. Journal of Educational Psychology, 8, 323-332.